Intellectica, 1991/1, 11, pp. 193-217

Pierre JACOB

INTERPRÉTATION, INFÉRENCE ET RATIONALITÉ *

Aussi vague que soit le sens du mot "rationnel", la propriété exprimée par ce mot s'applique censément à un système d'inférence, c'est-à-dire à un système qui forme et transforme des croyances. Il existe à ma connaissance, dans la littérature philosophique, trois arguments destinés à montrer que les processus cognitifs inférentiels des membres de l'espèce humaine doivent être crédités de *rationalité* ou qu'il existe des limites de principe à l'irrationalité qu'on peut censément leur attribuer. Quoi qu'on entende exactement par "rationalité", la rationalité que les philosophes sont enclins à prêter (comme Cohen 1981, 1986) ou à refuser (comme Stich 1990) aux processus cognitifs est une propriété compatible avec une pluralité de théories ou de modèles psychologiques employés pour expliquer les résultats expérimentaux — comme la *logique mentale* (défendue notamment par Braine 1990, Politzer ce numéro), les *schémas pragmatiques* (défendus notamment par Cheng & Holyoak 1985, Holland, Holyoak, Nisbett & Thagard 1986 ou Girotto ce numéro), les *modèles mentaux* (de Johnson-Laird 1983), et la dualité entre les *processus heuristiques* et les *processus analytiques* (défendue par Evans 1989).

Un seul de ces trois arguments philosophiques me paraît convaincant. Cet argument concerne les conditions d'attribution à un individu d'états

* Je remercie le relecteur anonyme de ses remarques pénétrantes.

intentionnels comme les croyances (et autres attitudes propositionnelles). Le mot "interprétation" dans mon titre n'a d'autre ambition que de faire référence à ce processus d'attribution d'intentionalité. Je voudrais faire valoir des raisons de ne pas souscrire aux deux autres arguments. Ensuite, je m'interrogerai pour savoir si le fait d'accepter l'argument pour limiter l'attribution d'irrationalité peut avoir une incidence plausible sur l'interprétation des résultats expérimentaux recueillis par les psychologues du raisonnement.

1. L'argument pour la rationalité par l'évolution

Il est un schéma d'argument plausible et plus ou moins répandu dans les esprits qui conclut à la rationalité des capacités inférentielles humaines à partir de la théorie de l'évolution par sélection naturelle : un système d'inférence qui a survécu aux épreuves de l'évolution par sélection naturelle doit être rationnel. On supposera — ce qui est naturel — que le fait pour un individu d'avoir un système d'inférence particulier est l'une de ses propriétés phénotypiques. Ce que les évolutionnistes nomment la *fitness*[1] d'un trait phénotypique s'analyse comme la contribution de ce trait à la *fitness* générale de l'organisme qui détient ce trait. On analyse par conséquent la *fitness* d'un système d'inférence par le détour de la contribution faite par ce système d'inférence à la *fitness* globale de l'organisme. Pour les besoins de cet argument, on identifiera la propriété exprimée par le mot "rationnel" à la *fiabilité* d'un système de formation de croyance, c'est-à-dire à la propriété qu'a un tel système d'engendrer plus de croyances vraies que de croyances fausses dans l'ensemble des croyances engendrées[2].

L'argument peut se représenter comme suit :

---

1. Sur l'emploi en français du mot anglais "fitness", cf. Gayon (1989 : 207-208).
2. Pour une discussion des conceptions "fiabilistes" de la connaissance et de la justification des croyances, cf. notamment Goldman (1986).

(1) L'évolution façonne des organismes dont les traits phénotypiques confèrent aux organismes une *fitness* optimale (disons pour aller vite que les traits sélectionnés ont une *fitness* optimale).

(2) Un système cognitif d'inférence dont la *fitness* est optimale est rationnel.

De (1) et (2) on conclut (3) :

(3) Les systèmes cognitifs inférentiels façonnés par l'évolution sont rationnels.

J'ai récemment été convaincu par la critique de cet argument effectuée par le philosophe Steve Stich (1990) et selon lequel il repose sur trop de propositions douteuses pour être acceptable.

Commençons par (2). (2) tire sa plausibilité de la conjonction de (2a) et (2b) :

(2a) La rationalité d'un système d'inférence est sa *fiabilité* : de deux systèmes S_1 et S_2, S_1 est plus *fiable* que S_2 si le pourcentage de croyances vraies (dans l'ensemble total des croyances) formées par S_1 est supérieur au pourcentage de croyances vraies formées par S_2.

(2b) De deux systèmes inférentiels S_1 et S_2, si la *fitness* de S_1 est supérieure à celle de S_2, alors S_1 est plus fiable que S_2 (au sens précisé en (2a)).

La faiblesse de (2b) tient à l'assimilation entre *fitness* d'un système inférentiel et fiabilité. Or pour évaluer la *fitness* d'un système inférentiel (la contribution du système inférentiel à la *fitness* de l'individu), il faut pondérer sa fiabilité par sa capacité à exploiter des ressources finies : sa

durée de vie, son énergie, sa mémoire. Toutes choses égales, on peut supposer que le pourcentage de croyances vraies dans l'ensemble des croyances est un élément de la *fitness* globale d'un système inférentiel. Mais un système enclin à inférer très vite et économiquement beaucoup de *fau*x *positifs* (du type "Il y a un tigre devant moi" en l'absence d'un tigre) et jamais de *fau*x *négatifs* (du type "Il n'y a pas de tigre devant moi" en présence d'un tigre) peut faire une contribution à la *fitness* globale de l'individu supérieure à un système qui engendre plus de vérités en prenant plus de temps et en gaspillant plus de ressources en énergie et en mémoire. La fausseté de (2b) prive (2) d'une grande part de sa plausibilité (même si on identifie rationalité et fiabilité).

Ce qui confère sa plausibilité à (1) sont les quatre propositions suivantes :

(1a) Tous les traits phénotypiques pertinents d'un individu sont façonnés par l'évolution.

(1b) L'évolution est entièrement commandée par la sélection naturelle.

(1c) Tout trait phénotypique ayant une *fitness* optimale est codé génétiquement.

(1d) Si elle a le choix, alors la sélection choisit toujours les gènes codant le trait ayant une *fitness* optimale.

(1c) et (1d) sont réfutées par les phénomènes dits de *pléiotropie* dont (i) l'*albinisme* et (ii) l'*anémie falciforme* sont des exemples : (i) le gène codant pour la fourrure blanche des ours polaires est aussi responsable d'une mauvaise vision ; (ii) les individus homozygotes souffrent de la maladie mais les individus hétérozygotes sont particulièrement résistants à la malaria. Le phénomène de la *dérive génétique* jette le doute sur (1b).

Considérons la capacité de parler le français (ou la connaissance du français) comme une propriété phénotypique d'un individu. Elle est la

source des intuitions de grammaticalité d'un francophone. Mais cette propriété phénotypique n'est pas directement façonnée par l'évolution — même si la capacité d'ordre supérieur à apprendre n'importe quelle langue humaine (la grammaire universelle au sens de Chomsky) est peut-être façonnée par l'évolution. La grammaticalité est une propriété des phrases d'une langue (disons du français). La connaissance tacite d'une langue (du français), non la capacité d'ordre supérieur à apprendre une langue quelconque, est la source des intuitions de grammaticalité d'un francophone.

Si on admet l'analogie entre aptitude inférentielle et aptitude linguistique, l'aptitude inférentielle d'un individu est l'une de ses propriétés phénotypiques. On peut sensément la qualifier de rationnelle (ou non). Mais ce qui a pu être sélectionné par l'évolution n'est pas, contrairement à ce que laissent entendre Cosmides (1989) et Cosmides et Tooby (1987), le système inférentiel d'un individu ; c'est la capacité d'ordre supérieur de l'individu à acquérir un système inférentiel. Cette capacité d'ordre supérieur n'est elle-même ni rationnelle ni irrationnelle et elle n'est pas la source directe des intuitions inférentielles de l'individu.

L’hypothèse de la sélection par l’évolution d’une aptitude à acquérir un système d’inférence particulier a à mes yeux l’avantage d’être compatible avec la théorie de l’évolution par sélection naturelle sans s’exposer aux critiques de Stich. Elle n’en doit pas moins être jugée empiriquement et non seulement conceptuellement. Pour au moins deux raisons, l’analogie entre aptitude à raisonner et aptitude à parler devrait peut-être être tempérée.

Premièrement, comme l’a fait remarquer Fodor (1975), si tant est que l’enfant qui apprend sa langue dispose de la connaissance tacite de ce que Chomsky nomme la grammaire universelle, il doit aussi disposer d’une capacité de raisonnement. Cette capacité de raisonnement, grâce à laquelle (en conjonction avec la grammaire universelle) l’enfant conjecture la grammaire de sa langue maternelle à partir de sa perception des données

linguistiques primaires, doit être un système d'inférence particulier, non une capacité d'ordre supérieur à acquérir un système d'inférence.

Mais logiquement (ou conceptuellement), l'enfant pourrait avoir acquis ce système d'inférence, sans que cela soit incompatible avec la thèse selon laquelle seule la capacité d'ordre supérieur à acquérir un système d'inférence, non un système d'inférence particulier, a été sélectionnée par l'évolution naturelle. Sans doute, dans le cadre des hypothèses de Chomsky et Fodor sur l'apprentissage du langage, l'enfant doit-il déjà posséder un système d'inférence particulier pour acquérir son savoir grammatical. La précocité de l'acquisition par l'enfant de la connaissance tacite de la grammaire de sa langue maternelle est-elle incompatible avec le fait qu'il ait acquis préalablement un système d'inférence particulier ? Seule la recherche empirique permettra de répondre.

Deuxièmement, c'est un fait que les tenants de l'hypothèse empirique de la logique mentale (cf. Braine 1990) supposent que tout enfant humain est équipé de la connaissance tacite d'un système d'inférence particulier, non de l'aptitude abstraite à acquérir un système d'inférence. Dans l'hypothèse où les tenants de la logique mentale ont raison d'attribuer à tous les membres de l'espèce humaine, dans toutes les cultures et à toutes les époques, la connaissance tacite d'un seul et même système d'inférence, l'analogie entre aptitude à raisonner et aptitude à parler n'est-elle pas réfutée par la diversité attestée des langues parlées par les humains adultes sur la surface du globe ? Si tous les humains adultes ont la connaissance tacite du même système inférentiel et s'ils ont la connaissance tacite de différentes grammaires de différentes langues naturelles, n'est-ce pas la preuve de la fausseté de l'analogie entre aptitude inférentielle et aptitude linguistique ?

La compatibilité de l'hypothèse de la sélection d'une aptitude de second ordre à acquérir un système d'inférence avec diverses approches retenues en psychologie expérimentale est à mes yeux un avantage. En outre, je ne crois pas qu'elle soit incompatible avec l'hypothèse d'une logique

mentale. En supposant vraie l’attribution à tous les humains adultes de la connaissance tacite d’un seul système inférentiel, je répondrais simplement — en défense de l’analogie — que peut-être l’environnement inférentiel se distingue-t-il de l’environnement linguistique par sa plus grande homogénéité. Les données ou stimuli linguistiques primaires mis à la disposition du nouveau-né humain par les membres de sa communauté diffèrent selon les communautés. En raison de l’homogénéité des stimuli linguistiques au sein d’une seule et même communauté, et en dépit de la diversité de leurs biographies, les enfants d’une seule et même communauté linguistique convergent sur la connaissance tacite d’une seule et même grammaire. Peut-être en dépit de la diversité des domaines dans lesquels elles s’exercent, les tâches d’apprentissage auxquelles sont confrontés les nouveaux-nés humains ont-elles une structure suffisamment uniforme pour assurer la convergence de leur aptitude de second ordre à acquérir un système inférentiel sur la connaissance tacite d’un seul et même système inférentiel.

2. Le deuxième argument pour la rationalité par l'équilibre réflexif et la distinction compétence/performance (L.J. Cohen)

Premièrement, Cohen (1981) admet le concept d'*équilibre réflexif* élaboré par Goodman (1955) pour caractériser la justification et la rationalité d'un système d'inférence. Le processus par lequel un logicien est conduit à tenir un système d'inférence (démonstrative ou non démonstrative) pour justifié est un processus d'ajustement (de va et vient) entre les règles d'inférence valides qu'il admet et les inférences particulières qu'il juge correctes. Il a donc des *intuitions* sur ce qu'est une inférence correcte.

(4) Le système d'inférence d'un logicien est justifié/rationnel s'il est en équilibre réflexif.

Pour que (4) ait une pertinence pour la *psychologie* du raisonnement (et non seulement pour la philosophie ou l'épistémologie de la logique), il faut pouvoir remplacer dans (4) "le système d'inférence d'un logicien" par "le système d'inférence d'un être humain profane" pour former (5) :

(5) Le système d'inférence d'un être humain profane est justifié s'il est en équilibre réflexif.

Pour justifier la transition de (4) à (5), Cohen fait un détour par la *linguistique générative*. Il fait en effet valoir que le concept d'équilibre réflexif peut servir à caractériser la démarche scientifique de la linguistique générative : le linguiste admet une règle grammaticale si elle est corroborée par les jugements intuitifs des locuteurs. Il la rejette si elle est infirmée par ces jugements. A l'occasion, il *répudie* un jugement intuitif qui entre en conflit avec des règles bien établies. Si le concept d'équilibre réflexif s'applique à la démarche de la linguistique générative, réciproquement la méthodologie de la grammaire générative ne vaut-elle pas pour la psychologie du raisonnement ? Je caractériserai cette méthodologie par (6a) et (6b) :

(6a) En linguistique, (i) les *intuitions* syntaxiques et sémantiques des locuteurs ordinaires sont tenues pour l'effet de leur *compétence* grammaticale et (ii) elles sont les seules autorités en matière de grammaticalité.

(6b) En linguistique, selon la fameuse distinction *compétence*/*performance*, un *jugement* d'acceptabilité peut être tenu pour l'effet conjugué de la *compétence* grammaticale (source des *intuitions*) et d'autres *compartiments cognitifs* (perception, mémoire, attention).

Nul doute que l'analogie entre la psychologie du raisonnement et la grammaire générative est importante pour la méthodologie de la

psychologie du raisonnement : la distinction compétence/performance de (6b) suggère que le psychologue du raisonnement ne peut pas automatiquement conclure d'une erreur dans une tâche de raisonnement à un défaut de la *compétence inférentielle* du sujet. S'il s'abstient d'imputer une erreur de raisonnement à un défaut de la compétence inférentielle du sujet, il lui incombe alors de fournir une explication concurrente et supérieure des faits expérimentaux, ainsi que le fait Politzer (ce numéro) lorsqu'il explique l'*erreur de conjonction* en se servant d'une hypothèse pragmatique sur la compréhension d'un énoncé contenant le mot français *et*. Autrement dit, Cohen semble justifié à transformer (6b) en (7b) :

> (7b) En psychologie du raisonnement, selon la fameuse distinction compétence/performance, un jugement inférentiel peut être tenu pour l'effet conjugué de la compétence inférentielle et d'autres compartiments cognitifs (perception, communication, mémoire, attention).

Cohen ne serait-il donc pas justifié à remplacer dans (6a)(i) "intuitions syntaxiques et sémantiques des locuteurs ordinaires" par "intuitions inférentielles des sujets", "compétence grammaticale" par "compétence inférentielle" et dans (6a)(ii) "grammaticalité" par "validité inférentielle" pour former (7a) ?

> (7a) En psychologie du raisonnement, (i) les intuitions inférentielles des sujets sont tenues pour l'effet de leur compétence inférentielle et (ii) elles sont tenues pour les seules autorités en matière de validité inférentielle.

Je voudrais apporter des raisons de douter de cette transformation.

La *linguistique* dans le programme de Chomsky est, comme l'a souligné Chomsky, un chapitre de la *psychologie*. Une grammaire générative du français est une hypothèse psychologique sur l'équipement cognitif des francophones. Ecrire une grammaire générative du français, c'est édifier

une théorie *explicite* d'un système de règles que *connaissent tacitement* les francophones. Dans ce programme, le linguiste cherche à définir la classe des langues humaines possibles. En caractérisant la grammaire générative d'une langue comme le français, il contribue à expliquer comment un enfant francophone acquiert la grammaire du français à partir des données linguistiques primaires mises à sa disposition par les membres de sa communauté. Pour trancher entre des hypothèses grammaticales rivales, toutes sortes de données psychologiques, neuroanatomiques ou électrophysiologiques peuvent être pertinentes.

On peut — comme Katz (1981) — contester ce programme et lui préférer une interprétation platonicienne de la linguistique. On distingue l'arithmétique et la psychologie de l'arithmétique comme on distingue la démonstration mathématique explicite d'une vérité arithmétique de l'exploration psychologique des connaissances tacites des vérités et/ou des faussetés arithmétiques dont les sujets humains non mathématiciens ont la connaissance tacite. Toutes les vérités arithmétiques ne se réduisent pas à celles dont les sujets non mathématiciens ont la connaissance tacite. On peut aussi distinguer entre la linguistique mathématique qui étudie les grammaires de tous les langages possibles et l'étude psychologique des grammaires des langues humaines possibles. Un membre de l'espèce humaine ne peut avoir la connaissance tacite que de l'une (ou plus) des grammaires des langues humaines — un sous-ensemble de l'ensemble des grammaires de tous les langages possibles. Pertinentes pour l'étude psychologique des grammaires des langues humaines possibles, les données psychologiques ou neuroanatomiques seraient dénuées de pertinence pour la linguistique mathématique[3].

Pas plus que la linguistique mathématique n'est un chapitre de la psychologie du langage, la *logique déductive* n'est un chapitre de la psychologie du raisonnement démonstratif et le calcul des probabilités n'est un chapitre de la psychologie du raisonnement dans l'incertitude.

---

[3]. Telle est apparemment la position défendue par Katz (1981 : 76-93).

Réciproquement, prouver un théorème en logique ou en calcul des probabilités, ce n'est pas faire une contribution à la psychologie du raisonnement. La démonstration d'un théorème logique n'est pas la codification explicite d'un enchaînement déductif entre des axiomes et un théorème dont un sujet non logicien aurait la connaissance tacite. Une théorie logique est aussi indépendante des intuitions inférentielles des sujets non logiciens qu'une théorie physique est indépendante des intuitions physiques des sujets non physiciens. Elle joue un rôle *normatif* vis-à-vis du système d'inférence et des intuitions d'un sujet non spécialiste. La tâche d'un psychologue du raisonnement démonstratif est d'étudier la *connaissance tacite* qu'a un sujet non spécialiste des règles logiques valides ou invalides.

Pertinentes pour la *psychologie* du raisonnement, les données de la neuroanatomie, de l'électrophysiologie et la connaissance des compartiments cognitifs comme la mémoire, la perception ou l'attention sont *dénuées de pertinence* pour la logique ou le calcul des probabilités. La logique et le calcul des probabilités sont indépendants des limitations cognitives propres à l'espèce humaine. Sans doute les intuitions syntaxiques et sémantiques d'un locuteur sont-elles les seules sources (ou autorités) en matière de grammaticalité. Mais si, en matière de validité inférentielle, le psychologue du raisonnement conférait aux intuitions inférentielles des sujets non logiciens la même autorité que celle que le linguiste confère en matière de grammaticalité aux intuitions syntaxiques et sémantiques des locuteurs ordinaires, alors il devrait renoncer à étudier les *erreurs* de raisonnement. Du moins se priverait-il de la possibilité de décrire certaines inférences comme des *erreurs* de raisonnement.

3. L'argument pour la rationalité par l'attribution d'intentionalité (Quine, Davidson, Dennett)

L'argument est simple :

(8) Un système d'inférence est un système qui modifie des croyances (par introduction de nouvelles croyances, élimination d'anciennes croyances, etc.).

(9) On ne peut attribuer à un individu de croyances (et d'autres attitudes propositionnelles, désirs, intentions, etc.) qu'en créditant l'individu de rationalité.

De (8) et (9), on conclut à (10) :

(10) Pour posséder un système d'inférence, un individu doit être rationnel.

Certes, parce que les erreurs de raisonnement et les phénomènes de dissonance cognitive sont si bien attestés, il est tentant de rejeter (9). Pour rejeter (9), deux stratégies semblent se présenter.

La première stratégie consisterait à faire valoir que dans (9) sont confondues deux propriétés distinctes d'un système intentionnel (ou système d'attitudes propositionnelles) : la *rationalité* et l'*intelligibilité*. Il se pourrait que la seconde mais non la première propriété soit une condition nécessaire pour l'attribution de croyances. Pour justifier cette distinction, on peut sans doute invoquer des cas réels ou imaginaires de ce que les anglophones nomment *wishful thinking* où un individu prend ses désirs pour la réalité. Typiquement, un individu qui prend ses désirs pour la réalité forme la croyance que *non p* lorsque, par exemple, ses mécanismes perceptuels lui donnent des raisons de former la croyance que *p* ; mais il a des raisons indépendantes de croire que *p* est une condition suffisante de *q* et il abhorre *q*. Un tel individu serait donc enclin à croire

que *non p* et à exprimer verbalement la proposition que *non p* lorsque les membres de sa communauté seraient enclins à croire et à exprimer la proposition que *p*. Quoiqu'un individu qui prend ses désirs pour la réalité ne soit pas rationnel, il n'en est pas moins intelligible, comme le soutient Levin (1988) : on peut censément lui attribuer des croyances (comme le prouve la reconstruction précédente).

A cela, un partisan de (9) pourrait répondre deux choses. (A) Premièrement, il pourrait concéder que l'individu qui prend ses désirs pour des réalités et forme incorrectement la croyance que *non p* n'est pas rationnel et faire valoir qu'on ne saurait cependant le créditer d'irrationalité sans conjointement créditer de rationalité les autres membres de sa communauté qui croient correctement que *p*. Autrement dit, l'attribution d'irrationalité à un individu présuppose (ou s'accomplit sur fond de) l'attribution de rationalité aux membres de sa communauté.

A l’affirmation selon laquelle l’attribution d’irrationalité à un individu présuppose la rationalité des membres de sa communauté, le partisan de la distinction entre rationalité et intelligibilité ne pourrait-il pas objecter que ma propre critique de la transformation de (6a) en (7a) (dont, selon moi, Cohen a besoin, cf. sec. 2) démontre qu’un psychologue expérimental (comme Tversky et Kahneman) peut découvrir empiriquement que, dans leur quasi-totalité, les membres d’une communauté font preuve d’irrationalité dans leur maniement du calcul des probabilités ? A cette objection, je répondrais deux choses.

Premièrement, je ne peux rendre intelligible le comportement précité de l’individu qui prend ses désirs pour la réalité qu’en comparant les mécanismes de fixation de sa croyance perceptuelle que *non p* aux mécanismes de fixation de la croyance perceptuelle que *p* des autres membres de sa communauté. Je ne peux m’abstenir de faire référence aux mécanismes de fixation des croyances perceptuelles des autres membres de la communauté qui servent à caractériser la fixation rationnelle des

croyances perceptuelles pour éclairer par des raisons le comportement de l’individu qui prend ses désirs pour la réalité.

Deuxièmement, le cas de la fixation des croyances perceptuelles est différent du cas de la fixation des croyances étudiées par Tversky et Kahneman comme la croyance selon laquelle la probabilité d’une conjonction est ou n’est pas supérieure à la probabilité de chaque membre de la conjonction. Je serais tenté de soutenir que, dans ce dernier cas, à la différence du cas des croyances de l’individu qui prend ses désirs pour la réalité, c’est le manuel d’utilisation du calcul des probabilités rédigé par un mathématicien qui sert à caractériser la rationalité des croyances des membres d’une communauté, non l’opinion la plus fréquemment répandue dans la communauté.

(B) Deuxièmement, il pourrait contester l'attribution d'irrationalité à l'individu qui prend ses désirs pour des réalités en faisant valoir qu'on ne peut pas distinguer l'intelligibilité de la rationalité. On ne peut en effet rendre intelligible l'enchaînement des croyances de l'individu sans les relier par des *raisons*. Si on accorde un degré suffisamment élevé à l'aversion de l'individu pour l'état de choses *q* et si on crédite l'individu de la croyance que *p* est une condition suffisante de *q*, alors on comprend les *raisons* qu'a l'individu de modifier sa croyance que *p* en croyance que *non p*. Notez qu'il est difficile de rendre l'individu intelligible sans lui attribuer préalablement la croyance perceptuelle que *p* et sans supposer un mécanisme de modification de la croyance que *p* en croyance que *non p*. Je tiens (A) et (B) pour une défense satisfaisante de (9) et de l'argument par l'attribution d'intentionalité.

Pour illustrer la seconde stratégie, imaginons un cas d'erreur *écologiquement* plausible (discuté part Stich 1982 et Dennett 1982). J'achète un verre de jus d'orange à un stand tenu par une fillette de douze ans. Le verre vaut 2,40F. Je lui donne une pièce de 5F. Elle me rend deux pièces de 1F et deux pièces de vingt centimes. Je lui signale qu'elle s'est trompée et qu'elle me doit encore une pièce de vingt centimes. Elle rougit,

s'excuse et me donne une pièce de vingt centimes supplémentaires. N'avons-nous pas un contre-exemple à la thèse selon laquelle nous devons créditer de rationalité le système d'inférence de la fillette ? Pour deux raisons, je crois avec Dennett que non.

Premièrement, le partisan de la rationalité doit concéder qu'en créditant le système inférentiel de la fillette de rationalité, il ne pouvait pas prévoir l'erreur de *calcul* de la fillette. Mais il peut faire valoir en revanche que, compte tenu de son erreur, en qualifiant son système d'inférence de rationnel et en supposant qu'elle n'avait nulle intention de me voler de 20 centimes, il peut alors prévoir le comportement de la fillette consécutif à son erreur de calcul. Devenue consciente de son erreur, elle manifestera un certain embarras (elle a rougi) : réciproquement, si la rationalité ne lui était pas imputée, son embarras deviendrait énigmatique.

Deuxièmement, son adversaire nie que l'hypothèse de rationalité soit une condition nécessaire d'attribution de croyances. Le partisan de la rationalité demandera alors à son adversaire d'établir la liste des croyances de la fillette telles que son erreur devienne intelligible. Supposons qu'elle croie :

(i) que je lui ai donné 5F ;
(ii) que le verre de jus coûte 2,40F ;
(iii) qu'une pièce de 5F vaut 5F ;
(iv) que deux pièces de 1F valent 2F ;
(v) qu'une pièce de vingt centimes vaut vingt centimes ;
(vi) que deux pièces de vingt centimes valent quarante centimes ;
(vii) qu'elle m'a rendu deux pièces de 1F et deux pièces de vingt centimes ;
(viii) que 5 - 2,40 = 2,60 ;
(ix) que 1 + 1 + 0,20 + 0,20 = 2,40 ;
(x) que 2,60 > 2,40 ;

et aussi :

(xi) qu'elle m'a rendu la monnaie correctement.

Comment concilier (xi) avec (i)-(x) ? Tel est le défi lancé par le partisan de l'hypothèse de rationalité à son adversaire. Une infraction à la rationalité dans un système d'inférence défie l'attribution de croyances. Une solution évidente serait de nier (vii) et de soutenir que la fillette croit incorrectement qu'elle m'a rendu deux pièces de un franc et *trois* pièces de vingt centimes : elle a formé incorrectement la croyance tactile qu'elle extrayait de son porte-monnaie et me tendait deux pièces de un franc et trois pièces de vingt centimes. Mais dans ce cas, la fillette n'est nullement irrationnelle et n'est pas un contre-exemple à l'hypothèse de la rationalité : elle a simplement une croyance perceptuelle *fausse* et me répondra vraisemblablement qu'elle croyait m'avoir rendu trois pièces de vingt centimes.

J’ai exploité le cas de l’individu qui prend ses désirs pour la réalité et le cas cas de la fillette qui commet une erreur de calcul pour faire valoir que l’adversaire de l’argument pour la rationalité par l’attribution d’intentionalité ne peut pas se prévaloir facilement d’une distinction entre rationalité et intelligibilité. Le critique pourrait m’objecter qu’en donnant des *raisons* du comportement de l’individu qui prend ses désirs pour la réalité, j’ai offert un embryon d’explication causale de son comportement ; je n’ai pas supposé qu’il était rationnel. Compte tenu de son aversion pour l’état des choses *q*, je ne suis pas convaincu que l’individu qui prend ses désirs pour la réalité soit moins rationnel que la plupart d’entre nous. La fillette qui commet une erreur de calcul représente un défi à qui voudrait prétendre qu’elle est irrationnelle.

4. L'argument n° 3 pour la rationalité peut-il avoir une incidence sur l'interprétation des résultats expérimentaux recueillis par les psychologues du raisonnement ?

Comme l'a souligné Davidson, l'hypothèse de la rationalité dans l'attribution d'états intentionnels est un principe *a priori* et non une hypothèse réfutable — ce n'est pas une "option" :

> ... si nous voulons attribuer de manière intelligible des attitudes et des croyances ou décrire utilement des mouvements comme un comportement, nous devons nous engager à trouver, dans le comportement, les croyances et les désirs un large degré de rationalité (Davidson 1974a : 237).

> La méthode ne consiste pas — car c'est impossible — à éliminer les désaccords ; son objectif est de rendre possibles des désaccords intelligibles et cela repose entièrement sur certains accords de base... Puisque la charité n'est pas une option mais la condition de formation d'une théorie, il est absurde de suggérer que nous pourrions commettre une massive erreur en l'admettant... Nous sommes contraints à la charité ; que nous l'aimions ou non, si nous voulons comprendre autrui, nous devons tenir la plupart de ses croyances pour correctes (Davidson 1974b : 197).

Ne serait-il pas surprenant sinon suspect qu'un principe *a priori* ait une incidence sur les recherches expérimentales ? Je veux suggérer justement que la conclusion de l'argument pour la rationalité par l'attribution d'intentionalité (que j'admets) ne devrait avoir *aucune* incidence sur les recherches expérimentales sur le raisonnement. Ma stratégie est de faire valoir que la prémisse (9) ne concerne pas les recherches expérimentales sur le raisonnement, qu'il existe deux versions de (8) — une version radicale et une version modérée —, que (8) dans sa version radicale est

fausse, que (8) dans sa version modérée est vraie et que l'argument n° 3 est compatible avec la version modérée.

En empruntant un mot au vocabulaire de la philosophie de Kant, on peut qualifier pompeusement de “transcendantal” tant l’argument de Cohen par la distinction compétente/performance que l’argument de Quine-Davidson-Dennett par l’attribution d’intentionalité. Un argument transcendantal pour la rationalité d’un système inférentiel est un argument visant à établir qu’il est incohérent (ou conceptuellement absurde) de qualifier d’irrationnels les processus cognitifs des sujets étudiés par les expérimentateurs. J’ai rejeté le premier, ne suis-je pas incohérent en acceptant (contrairement à Stich 1990 : ch. 2) le second — quitte à faire valoir qu’il n’a pas d’incidence sur la psychologie expérimentale du raisonnement ? Je peux, je crois, sans risque d’incohérence, rejeter un argument transcendantal qui me paraît fallacieux, sans pour autant rejeter *tous* les arguments de cette catégorie.

De quoi parle la prémisse (9) ? (9) parle de ce que j'appellerai les *croyances authentiques de la psychologie naïve* et affirme que la rationalité est une condition nécessaire d'attribution de croyances authentiques à un individu. Comme l'ont remarqué les philosophes, les croyances authentiques sont des *attitudes propositionnelles* ; elles peuvent être factorisées en deux composantes : un contenu propositionnel et une attitude vis-à-vis de ce contenu. Dans les cas les plus simples, un individu a une croyance authentique (ou en est crédité) lorsqu'il tient pour vraie une proposition. Comme le soulignent Sperber et Wilson (1986), dans leur environnement cognitif écologiquement normal (ou valide), les humains forment régulièrement des croyances authentiques selon trois voies : par la perception, par la communication et par des inférences *spontanées*. Dans les processus d'inférence spontanée, les croyances authentiques ont deux caractéristiques — l'une concerne le contenu, l'autre l'attitude. D'une part les contenus propositionnels sont des représentations ou des descriptions *complètes* d'états de choses particuliers. D'autre part les contenus propositionnels qui servent d'objet de croyance authentique sont *complètement compris* ou *interprétés* par les individus.

Loin d'être spontanées et implicites comme par exemple les inférences accomplies dans la communication verbale et non verbale ordinaire, les inférences auxquelles donnent lieu les tâches étudiées par les psychologues du raisonnement requièrent un effort conscient et elles sont menées *explicitement*. Certaines tâches sont même méta-inférentielles comme le sont la tâche de sélection de Wason ou la tâche de découverte de la règle de formation d'une séquence de trois nombres entiers en ordre ascendant (cf. Politzer, ce numéro). Est inférentielle une tâche qui consiste à appliquer une règle. Pour réussir la tâche de sélection de Wason, les sujets ne peuvent se contenter d'appliquer une règle, ils doivent de surcroît fournir une méthode permettant de vérifier et réfuter une règle. Dans la seconde tâche, il est demandé aux sujets de découvrir une règle. Les gricéens orthodoxes ou hétérodoxes (comme Sperber et Wilson 1986) s'accordent à reconnaître que, dans la communication humaine, la tâche d'inférence implicite et spontanée qu'un destinataire doit accomplir est facilitée par l'intention de son partenaire d'être *pertinent*. Mais dans l'énoncé d'un problème logique à résoudre, les psychologues livrent délibérément des informations *non* pertinentes et le souci des sujets d'exploiter toutes les informations qui leur sont livrées (y compris des informations non pertinentes) interfère régulièrement avec leur aptitude inférentielle à résoudre le problème.

Ces tâches inférentielles menées *explicitement* ou *délibérément* et non *spontanément* modifient non des croyances *authentiques* mais des *proto*-croyances — ou *croyances**. Un élève de mathématiques qui suppose vraie une proposition *p* dans le but de démontrer par l'absurde que *p* est fausse, a à l'égard de *p* une attitude de proto-croyance ou croit* *p*. Quoique le contenu d'une croyance* ait certaines propriétés logiques, il n'a pas besoin d'être la représentation ou la description complète d'un état de choses particulier. Et le sujet ne comprend pas toujours complètement le contenu logique d'une croyance*. Lorsqu'un sujet croit* un certain contenu qu'il comprend partiellement, il effectue une *simulation cognitive* : il

suppose vraie (il fait semblant de croire vraie) une proposition ou un contenu logique incomplètement propositionnel pour en examiner les conséquences qu'il est capable de calculer.

Considérons la phrase française (11) :

(11) Elle l'a transporté de la gare à l'école.

Cette phrase a une structure sémantique (ou signification linguistique). Si vous ne savez ni à qui ni à quoi il est fait référence respectivement par *elle* et par le clitique *l'*, ni de quelle gare ni de quelle école il est question dans l'énoncé de (11), vous ne savez pas quelle proposition l'énoncé de (11) sert à exprimer explicitement. Si vous ignorez tout cela, vous ne savez pas quel est l'état de choses particulier décrit par l'énoncé de (11). Dans le cadre *écologique* dans lequel (11) sert normalement à un locuteur pour communiquer sa pensée à un destinataire, celui-ci complète automatiquement les informations incomplètes véhiculées par la structure sémantique de la phrase en extrayant des informations supplémentaires de sa connaissance du contexte. Grâce à cette combinaison d'informations (les unes sémantiques, les autres contextuelles), le destinataire forme la représentation (ou la compréhension) d'une *proposition singulière* (russellienne) qui est la description complète d'un état de choses particulier. A défaut d'avoir des raisons de soupçonner l'énonciateur de (11) de ne pas être un informateur *fiable*, le destinataire entretient automatiquement une *croyance authentique* vis-à-vis de l'information qui lui est communiquée — la description complète d'un état de choses singulier. Par défaut, il croit vraie une proposition singulière qu'il comprend.

Mais grâce à votre seule compréhension du sens de la phrase, sans rien savoir du contexte, vous pouvez aussi former une *proposition générale* truffée de quantificateurs existentiels et affirmant *qu'il existe* une gare indéterminée, une école indéterminée, un objet indéterminé et un être animé indéterminé et que l'objet indéterminé était dans la gare

indéterminée avant l'énonciation de (11) ; et que l'objet indéterminé est dans l'école indéterminée après l'énonciation de (11) ; et que le changement de lieu occupé par l'objet indéterminé a été occasionné par un déplacement de l'être animé indéterminé. Cette proposition *générale* n'est pas la proposition singulière exprimée par un énoncé de (11) et comprise par un destinataire dans un contexte écologiquement normal d'utilisation de la langue : elle ne décrit aucun état de choses particulier. Elle n'en a pas moins des propriétés *logiques* : si elle est vraie, alors il est faux par exemple qu'aucun objet n'ait jamais changé de lieu. Vous pouvez *croire** cette proposition *générale*, la supposer vraie et en examiner les conséquences générales.

C'est exactement ce à quoi se livrent les sujets des expériences psychologiques et les élèves de mathématiques. Soit le prénom *Linda* dans l'énoncé du problème qui donne lieu à l'*erreur de conjonction* rendue fameuse par Tversky et Kahneman. Lorsque le prénom *Linda* est utilisée par un locuteur dans un échange d'information "écologiquement valide" (une conversation), il permet normalement à l'auditeur d'identifier un référent. Le traitement d'un tel prénom contribue donc à la compréhension d'une proposition singulière. Cependant, dans l'énoncé qui donne lieu à l'erreur de conjonction, *Linda* est un nom propre purement descriptif qui fait référence à une jeune femme quelle qu'elle soit qui satisfait les propriétés mentionnées dans le descriptif élaboré par les psychologues. Lorsqu'un sujet traite la phrase *Linda est caissière dans une banque*, il construit une proposition existentiellement quantifiée et s'y rapporte par une croyance*, non par une croyance authentique.

Dans sa version radicale, (8) soutient qu'un système d'inférence modifie et ne modifie que des croyances authentiques (formées dans un contexte écologiquement valide). Nous devons rejeter la version radicale de (8) et admettre une version modérée de (8) qui dit que les inférences peuvent agir sur des croyances* ou des proto-croyances puisque celles-ci ont (comme les croyances authentiques) des propriétés logiques. Grâce à ces

propriétés logiques, les croyances* (comme les croyances authentiques) entretiennent les unes avec les autres des relations d'implication, de compatibililité ou d'incompatibilité logique. Mais à ces proto-croyances, (9) ne s'applique pas vraiment : car la question de la fixation du contenu propositionnel complet n'est pertinente que pour les croyances authentiques, non pour l'objet d'une simulation cognitive. D'autant moins que l'objet d'une simulation cognitive peut être partiellement compris ou interprété.

Je voudrais prévenir deux confusions possibles. Premièrement, je ne prétends nullement — ce qui serait absurde — que toutes les croyances authentiques de la psychologie naïve sont des attitudes spontanément formées à l'égard de propositions *singulières* ou qu'aucune croyance authentique n'est une attitude à l'égard d'une proposition *générale*. J'ai choisi l'exemple de la proposition singlière exprimable par l'énoncé de la phrase (11) pour faire un contraste entre deux processus d'interprétation : un processus spontané écologiquement valide et un processus artificiel, délibéré ou non spontané. Le processus écologiquement valide donne lieu à la compréhension d'une proposition singulière. Le processus artificiel d'interprétation du sens de la phrase privé des informations contextuelles qui font partie du processus écologiquement valide donne lieu à la compréhension d'une proposition générale. A la différence du premier processus qui aboutit à la formation d'une croyance authentique, le second processus aboutit à la formation d'une croyance* ou proto-croyance.

Le même contraste peut être mis en évidence par référence à l'interprétation de l'énoncé de phrases contenant un quantificateur. Soit (12) :

(12) Aucun enfant n'est venu.

Dans un contexte écologiquement valide, non seulement la proposition exprimée par (12) contient une référence tacite à une paire de lieux <x, y> (dont l'un y est occupé par le locuteur), mais le quantificateur "aucun" est

automatiquement interprété comme un quantificateur *restreint* : l’énoncé fait implicitement référence au domaine par rapport auquel le quantificateur restreint doit être interprété — par exemple, le domaine des enfants d’une classe particulière dans une école particulière. Autrement dit, dans un contexte écologiquement valide, l’énoncé de (12) sert à exprimer ou communiquer la proposition générale qu’aucun des membres de l’ensemble délimité des enfants d’une classe particulière d’une école particulière ne s’est rendu du lieu x au lieu y (occupé par le locuteur au moment de l’énonciation de (12)). Cette restriction de l’interprétation du quantificateur est automatiquement opérée par l’auditeur qui connaît le contexte. Mais du point de vue de la logique standard du premier ordre, cette phrase sera dite exprimer la proposition qu’aucun enfant dans l’univers n’est venu, qui est la négation de la proposition que quelques enfants sont venus. Aucun des enfants susceptibles de servir de contre-exemple à l’interprétation purement logique de l’énoncé de (12) ne sera retenu contre l’interprétation écologiquement valide. Autrement dit, le fait qu’un enfant qui ne fait pas partie du domaine de restriction du quantificateur soit venu ne réfute pas la proposition obtenue par le processus d’interprétation écologiquement valide.

Deuxièmement, il peut paraître paradoxal de soutenir — comme je le fais — que l’argument pour la rationalité par l’attribution d’intentionalité s’applique à la formation et à la modification spontanées (non délibérées) des croyances authentiques de la psychologie naïve, et non aux processus explicites et délibérés de raisonnement par lesquels sont formées et modifiées des croyances* ou des proto-croyances qu’étudient les psychologues expérimentaux. Ce qui devrait atténuer l’impression de paradoxe est le fait suivant : qu’il s’agisse de la proposition respectivement exprimée par (11) ou par (12), le processus écologiquement valide d’interprétation de l’énoncé consiste à *enrichir* spontanément par des informations contextuelles la forme logique qui correspond au sens de la phrase énoncée. Dans (11), l’enrichissement

provoque la modification d'une proposition générale en proposition singulière et il inclut la référence à un domaine de quantification dans (12). Le processus de formation d'une proto-croyance ou d'une croyance* portant sur la contrepartie des propositions normalement communiquées par un énoncé de (11) ou de (12) consiste à bloquer ou prohiber ce processus d'enrichissement.

Dans l'argument par l'attribution d'intentionalité, la rationalité est prédiquée des processus de formation et de modification de croyances portant sur des objets logiques ayant subi ce processus d'enrichissement. J'ai essayé de faire valoir dans cette section que la rationalité au sens de Quine-Davidson-Dennet ne s'applique pas à des processus de formation et de modification de croyances* ou de proto-croyances ayant pour contenus des objets logiques amputés de cet enrichissement spontané.

Pierre JACOB
C.N.R.S. et École Polytechnique
Centre de Recherche en Épistémologie Appliquée
1, rue Descartes
75005 PARIS

Références

Braine, M.D.S. (1990) 'The "Natural Logic" Approach to Reasoning', in W.G. Overton (éd.) *Reasoning, Necessity, and Logic: Developmental Perspectives,* Hillsdale, N.J.: Erlbaum.

Cheng, P.W. & K.J. Holyoak (1985) "Pragmatic Reasoning Schemas", *Cognitive Psychology,* 17, 391-417.

Cohen, L.J. (1981) "Can Human Irrationality Be Experimentally Demonstrated?", *The Behavioral and Brain Sciences,* 4, 317-70.

Cohen, L.J. (1986) *The Dialogue of Reason, An Analysis of Analytical Philosophy,* Oxford: Clarendon Press.

Cosmides, L. (1989) "The Logic of Social Exchange: Has Natural Selection Shaped How Humans Reason? Studies with Wason Selection Task", *Cognition*, 31, 187-276.

Cosmides, L. & J. Tooby (1987) "From Evolution to Behavior: Evolutionary Psychology as the Missing Link", in J. Dupré (éd.) *The Latest on the Best,* MIT Press.

Davidson, D. (1974a) "Philosophy as Psychology", in D. Davidson (1980) *Essays on Actions and Events,* Oxford: Clarendon Press.

Davidson, D. (1974b) "On the Very Idea of a Conceptual Scheme", in D. Davidson (1984) *Inquiries into Truth and Interpretation,* Oxford: Clarendon Press.

Dennett, D. (1982) "Making Sense of Ourselves", in D. Dennett (1987) *The Intentional Stance,* Cambridge, Mass.: MIT Press.

Evans, J.St. (1989) *Bias in Human Reasoning. Causes and Consequences,* Hillsdale, N.J.: Erlbaum.

Fodor, J.A. (1975) *The Language of Thought,* New York: Crowell

Gayon, J. (1989) "Epistémologie du concept de sélection", in P. Jacob (éd.) *L'Age de la science,* II, "Epistémologie", 201-27.

Goldman, A. (1986) *Epistemology & Cognition,* Cambridge, Mass.: Harvard UP.

Goodman, N. (1955) *Fact, Fiction and Forecast,* New York: Bobbs Merrill.

Holland, J.H., K.J. Holyoak, R.E. Nisbett & P.R. Thagard (1986) *Induction: Processes of Inference, Learning and Discovery,* Cambridge, Mass.: MIT Press.

Johnson-Laird, P. (1983) *Mental Models,* Cambridge: Cambridge UP.

Katz, J.J. (1981) *Language & Other Abstract Objects,* Totowa, N.J.: Rowman & Littlefield.

Levin, J. (1988) "Must Reasons Be Rational?" *Philosophy of Science,* 55, 199-217.

Quine, W.V.O. (1960) *Word and Object,* Cambridge, Mass.: MIT Press.

Sperber, D. & D. Wilson (1986) *Relevance, Communcation and Cognition,* Cambridge, Mass.: Harvard UP.

Stich, S. (1982) "Dennett on Intentional Systems", in J.I. Biro & R.W. Shahan (éds.) *Mind, Brain and Function,* Brighton: Harvester Press.

Stich, S. (1990). *The Fragmentation of Reason,* Cambridge, Mass.: MIT Press.